Using a Human Disease Network for Augmenting Prior Knowledge about Diseases

Hossein Rahmani [1a], Hendrik Blockeel[b], Andreas Bender[c]

[a] *Faculty of Arts and Social Sciences, Universiteit Maastricht, PO Box 616 6200 MD, Maastricht, The Netherlands*

[b] *Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan 200A, 3001 Leuven, Belgium*

[c] *Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom*

Abstract

The cellular metabolism of a living organism is among the most complex systems that man is currently trying to understand. Part of it is described by so-called protein-protein interaction (PPI) networks, and much effort is spent on analyzing these networks. Recently, there has been much interest in predicting involvement of network nodes (in this case, proteins) in different diseases. Many approaches to this problem exist. We categorize the previous studies into Individual and Network approaches. While the Individual approach focuses on one specific disease without considering its relationship with other diseases, the Network approach considers also these relationships. In this paper, we construct a Human Disease Network (HDN), using a novel approach for discovering relationships among different diseases. We built the HDN for 20 different diseases based on functional and structural information available in the PPI network. We showed that the proposed HDN is biologically meaningful and is capable of augmenting the initial prior knowledge of different diseases by sharing information across highly-related diseases. Furthermore, comparing to previous Individual and Network approaches, our proposed HDN increases the accuracy of predictive models and discovers more and still informative relationships among different diseases, respectively.

Keywords: Protein-Protein Interaction (PPI) Network, Disease-related

[1]Corresponding author: Hossein.Rahmani@MaastrichtUniversity.nl

1. Introduction

In recent years, much effort has been invested in the construction of protein-protein interaction (PPI) networks [1]. Much can be learned from the analysis of such networks with respect to the metabolic and signalling processes present in an organism, and the knowledge gained can also be prospectively employed e.g., for the task of protein function prediction [2, 3, 4], identification of functional modules [5], interaction prediction [6, 7], identification of disease candidate genes [8, 9, 10, 11, 12, 13, 14, 15, 16, 17] and drug targets [18, 19], according to an analysis of the resulting network [20].

Wu et al. [16] present an excellent overview of multiple methods for detecting proteins involved in disease or cancer. Among the different methods discussed in [16], "guilt-by-proximity" methods are well known. Methods classified in this category are based on the assumption that genes that directly interact, or, more generally, lie close to each other in the network, are more likely to be involved in the same diseases (as argued by, e.g., Gandhi et al. [21]). The methods vary based on how they define proximity: Some methods consider only direct neighbors to be in the proximity (e.g., [15, 22]), some quantify proximity of two proteins using the length of the shortest-path between them, some compute a "Global Distance Measure" that also takes into account how many paths there are between the two proteins, and how long these are; an example is the approach by Chen et al. [23], who use a PageRank based model for this.

The methods discussed by Wu et al. [16] mostly rely on notions of proximity (to genes known to be disease-related) from the area of graph analysis. An entirely different type of approaches are those that rely on feature-based descriptions [17, 24, 25, 26]. There, each individual protein is described by means of a fixed set of features. Next, using machine learning methods, a model is learned that links some of these features to disease-relatedness.

In addition to methods which analyze each disease individually (Individual Approach), other work has considered also the relationships among human diseases (Network Approach) [27, 28, 29, 30, 31, 32]. Among these network based approaches, Goh et al.'s method [27] has a prominent role. This method connects two diseases in the network if there is at least one gene that is implicated in both. However, their approach is not capable of discovering relationships among diseases with no common disease-related genes.

In this paper, we use contextual information from PPI networks (Structural and Functional information), which is shown informative by previous studies [33, 34], as one possible way of solving the limitations of Goh et al.'s method. In the resulting network, two diseases even with no common disease-related genes may be connected to each other if they are found to be similar with respect to contextual information extracted from PPI networks. Our proposed network improves Goh et al.'s method with respect to coverage (more relationships among diseases) and accuracy. Once the Human Disease Network (HDN) is constructed, the information in the HDN can be used to obtain new predictions, identifying genes that may be involved in some disease but would not be detected using previous Individual and Network approaches.

We describe the proposed method in detail in Section 2. In Section 3, the method is applied to a concrete PPI network; using functional and structural information from the network, a disease network is constructed for 20 diseases. This network is next evaluated in terms of interpretability and knowledge augmentation, and compared to a disease network constructed using previous methods. Section 4 presents a case study, where the predictions for one particular disease are analyzed and interpreted in detail. Section 5 concludes.

2. Methods

2.1. Terminology and Symbols

We consider a PPI network as an undirected annotated graph $(P, E, \lambda_F, \lambda_D)$ where P is a set of proteins, $E \subseteq P \times P$ is a set of interactions between these proteins, and λ_F and λ_D are so-called annotation functions; for each protein p, λ_F and λ_D denote additional information we have about p. $\lambda_F(p)$ lists all the GO functions that are associated with p; we call it the function set (or function vector) of p, and denote it $FS(p)$. $\lambda_D(p)$ lists all the diseases that protein p is known to be involved in; we call it the disease list of p and denote it $dizList(p)$. If $D = \{d_1, d_2, \ldots, d_m\}$ is the list of m analyzed diseases in our paper, then $dizList_i(p) = 1$ if p is involved in d_i and 0 otherwise. We also define seed proteins $SP(d_i)$ as the set of proteins involved in disease d_i ($d_i \in dizList(p) \Leftrightarrow p \in SP(d_i)$).

2.2. Human Disease Network

We define a Human Disease Network (HDN) as a directed graph $HDN(D, R)$ where D is a set of diseases and $R \subseteq D \times D$ is a set of directed relationships

between these diseases. We build our proposed HDN as follows.

For each disease d_i, we learn a model that can predict, for any protein p, how likely p is involved in this disease. Next, we use this model to make predictions for all the seed proteins of a disease d_j. The higher these seed proteins score, on average, the stronger the link between d_i and d_j is considered to be.

Concretely, the model for a disease d_i is learned and used as follows:

1. let $testSet$ contain the seed proteins of all diseases except d_i.
2. let $trainSet$ contain all proteins not in $testSet$
3. We learn a predictive model M from $trainSet$, using the seed proteins of d_i as positive examples and all other proteins as negative examples. We then use M to predict for each protein in $testSet$ how likely it is involved in d_i (higher values meaning more likely). For randomized learners, we repeat this 10 times (otherwise just 1 time) and calculate for each $p \in testSet$ the average, denoted $APV(p)$.
4. For each disease $d_j \in D(j \neq i)$, we add a directed edge $d_i \rightarrow d_j$ to the HDN with a weight

$$weight(d_i \rightarrow d_j) = \frac{\sum_{p \in SP(d_j)} APV(p)}{|SP(d_j)|} \quad (1)$$

with $|SP(d_j)|$ the number of seed proteins of d_j.

This procedure is repeated for all diseases. The resulting HDN is a directed, fully connected network in which each node is a disease and each weighted edge shows a relationship between two diseases. A high weight for $d_i \rightarrow d_j$ expresses that proteins involved in d_j are, on average, likely to be involved also in d_i, according to the model built for d_i.

In order to focus on the most important relationships in the HDN, we prune the network by keeping only the highest-ranked edges.

There are many ways in which the predictive model M can be learned from the PPI network (step 3). In the following section, we discuss those that we have experimented with.

2.3. Prediction Methods

Many methods for predicting which proteins are involved in certain diseases have been proposed. We here distinguish three types: methods that use the *structural* information in the PPI network, methods that use *functional* information, and hybrid methods (which use both).

2.3.1. Structural Category: Random Walk based Method (ST-RW)

Berger et al. [34] assume that disease-related proteins fall closer on average to the seed proteins than they do on average to the rest of the network. They calculate the score of each protein p_j in the network based on Formula 2 and then, select high-scoring proteins as disease-related proteins.

$$score_s(p_j) = \frac{\frac{\sum_{i \in C'} T_{ij}}{|C'|} - \frac{\sum_{i \in C} T_{ij}}{|C|}}{\frac{\sum_i T_{ij}}{|C|+|C'|}} \tag{2}$$

In Formula 2, T_{ij} is the average number of steps a random walker takes to walk from a specified node i to another specified node j, C is the set of seed proteins and C' is the set of all other proteins in the network. In the rest of this paper, we refer to this method as *ST-RW*.

2.3.2. Structural Category: ANOVA based Method (ST-Anova)

Rahmani et al. [3] proposed a relevance measure for proteins that is inspired by analysis of variance (ANOVA), and showed that shortest-path distance to a relatively small number of proteins (selected according to the ANOVA-based measure) is informative for the task of function prediction in the PPI network. Since this method works well for function prediction, we consider it also for the task of predicting which proteins are involved in some disease.

Let D be the set of proteins labeled with disease diz, and $\bar{D}$ the set of proteins not labeled with it. Given a particular protein q, if proteins in D tend to have a shortest-path-distance to q that is very different from proteins in $\bar{D}$, then "shortest-path-distance to q" is an informative feature. We call this feature δ_q. How informative δ_q is, can be measured using standard analysis of variance. Let $E_S(\delta_q)$ and $Var_S(\delta_q)$ denote the sample mean and variance of δ_q in a set S, respectively; then

$$A_q = \frac{(E_D(\delta_q) - E_{\bar{D}}(\delta_q))^2}{Var_D(\delta_q) + Var_{\bar{D}}(\delta_q)} \tag{3}$$

indicates to what extent the shortest-path-distance to q correlates with being in D. A high A_q means that δ_q varies little within groups and/or much between groups, which indicates that δ_q has high predictive power for the group (i.e., for whether the protein is involved in the disease or not). Features δ_q can be ranked according to A_q, and the top-k features selected as actual features to be included in the description of all proteins.

A standard machine learning system can next be used to learn a model that from these informative features predicts the likelihood of involvement. We here use Naive Bayes [35]. This method estimates the conditional probability distribution for the variable to be predicted, given the feature values, and up to a constant factor, as follows:

$$p(C|F_1, \ldots, F_n) \sim p(C)p(F_1|C)p(F_2|C)\cdots p(F_n|C) \tag{4}$$

This estimation of the conditional probability distribution relies on conditional independence of the features given the target. This assumption is usually violated, but the method is quite robust to violations of the assumption [36], and works well in practice. Furthermore, several researchers found, in a similar context, that the features are more important than the actual machine learning method used [26].

The method of first selecting δ_q features using ANOVA, then applying Naive Bayes to obtain the predictive model, is henceforth referred to as *ST-Anova*.

2.3.3. Functional Category: Individual based Method (Func-Indiv)

This method uses the functional annotations of proteins. For each function, it determines how strongly the function correlates with disease-relatedness, using the standard χ^2 statistic as proposed by Liu et al. [37]:

$$\chi^2(f_i) = \frac{(ad - bc)^2 * (a + b + c + d)}{(a + b)(c + d)(b + d)(a + c)} \tag{5}$$

where $a = |D \cap P_i|$, $b = |D \cap \bar{P}_i|$, $c = |\bar{D} \cap P_i|$, and $d = |\bar{D} \cap \bar{P}_i|$, with P_i and $\bar{P}_i$ the set of proteins labeled and not labeled with function f_i, and D and $\bar{D}$ the set of disease-related and not disease-related proteins, respectively.

Next, it describes all proteins using the highest-scoring functions as features, and applies Naive Bayes to obtain a predictive model that uses these features as input. In the rest of this paper, we refer to this method as *Func-Indiv*.

2.3.4. Functional Category: Collaboration based Method (Func-Collab)

Selecting individual discriminative functions based on $\chi^2(f_i)$ does not consider the network topology and the way different functions interact with each other in the network. Rahmani et al. [38] showed that for the task of predicting cancer-related proteins, it is possible that a function f_i does not

correlate itself with cancer-involvement, but interaction of the same function with some function f_j does correlate with the former protein being involved in a cancer. Rahmani et al. [38] proposed a new way of calculating the χ^2 of the function pairs in the PPI network.

They select high-ranked collaborative function pairs and then, they describe the proteins based on the high-ranked function pairs. In the end, they applied the naive Bayes classifier for predicting the proteins involved in cancer. In the rest of this paper, we refer to this method as *Func-Collab*.

2.3.5. Hybrid Category: Integrating Functional and Structural Information

Structural-based and functional-based methods can be simply combined into hybrid methods as follows:

$$Hyb(p) = norm(ST(p)) + norm(Func(p)) \tag{6}$$

where $ST(p)$ and $Func(p)$ represent the disease-relatedness score of p using a *Structural* (*ST-RW* or *ST-Anova*) and a *Functional* (*Func-Indiv* or *Func-Collab*) method, respectively. In order to avoid that one category dominates, we normalize the disease-relatedness scores using

$$norm(x_i) = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{7}$$

where $min(x)$ and $max(x)$ return the minimum and maximum values taken over all values of x, respectively.

Combining any structural method with any functional method gives four hybrid methods, as shown in Table 1.

3. Empirical Results and Discussion

In this section, we discuss several aspects of the proposed method in more detail, and investigate them experimentally.

The dataset used for the experiment is described in Section 3.1.

Recall that our method relies on a predictive model M, constructed using one of the methods listed in Section 2.3. As the quality of the constructed HDN is likely to depend on the quality of this model, in Section 3.2 we try to determine experimentally which method is most reliable.

In Section 3.3, we show the HDN constructed by our method, and provide a biological interpretation; this is meant as an evaluation of how informative and interpretable the network is.

Our method constructs the HDN from predictions about involvement in a disease, rather than from confirmed data (the "seed proteins"). One may wonder whether constructing the HDN using only the confirmed data would work better. In fact, this procedure would correspond to the method proposed earlier by Goh et al. [27]. In Section 3.4, we briefly discuss earlier proposals for constructing human disease networks and experimentally compare our HDN to Goh et al.'s.

Finally, a more objective type of evaluation is: how useful is the HDN from the point of view of predicting involvement in diseases? Does using the network yield better predictions? In section 3.5, we augment the initial seed proteins of different diseases by sharing information across highly-related diseases. Then, we evaluate the use of augmented seed proteins for predicting disease-related proteins in two different ways.

3.1. Dataset

We applied our method for building the HDN to the PPI network used by Milenkovic et al. [24]. This dataset is the union of three human PPI datasets: HPRD [39], BIOGRID [40] and the dataset used by Radivojac et al. [41] and contains $47,303$ physical interactions among $10,282$ proteins. When we say "union", we mean that the new network contains all the nodes and edges (proteins and interactions) found in either of these networks. The aim of merging these three datasets was to obtain as complete a human PPI network as possible, i.e., a network that covers with its edges as many proteins in the human proteome as possible. Milenkovic et al. [24] provide details on the construction of the integrated network. The GO functions of proteins are extracted from [42]. Table 2 shows some basic statistical information about our annotated dataset.

Table 3 shows the list of 20 different diseases analyzed in Aerts et al., [43], in addition to their "Seed Count" and "Augmented Seed Count". Seed Count is the number of proteins initially listed as involved in the disease. Augmented Seed Count is the number of proteins listed as involved after applying our method, as discussed in detail in section 3.5.

3.2. Choosing a Prediction Method

Section 2.3 mentions a list of methods that could be used for building a predictive model used to construct the HDN. We have experimentally determined which of these methods gives the most reliable predictive model on our dataset. The methods from Section 2.3 rely on a feature selection

step, for which a number of features needs to be decided. Based on earlier work on the same dataset [33], we consistently choose 100 functional and 10 structural features. Having made this choice, we compare the methods using the following leave-one-out cross-validation procedure, proposed earlier by De Bie et al. [44].

For each disease d_i:

1. Randomly select 99 proteins that are not seed proteins for d_i
2. For each seed protein p of d_i:
   (a) let $testSet$ contain p and the 99 proteins
   (b) let $trainSet$ contain all other proteins
   (c) Learn a model M from $trainSet$, apply it to $testSet$, and check how high p ranks.
3. Repeat steps 1 to 2 ten times and calculate the average rank $AR(p)$ of each seed protein p.

The overall rank of d_i, for a given method, is then defined as

$$overallRank(d_i) = \frac{\sum_{p \in SP(d_i)} AR(p)}{|SP(d_i)|} \tag{8}$$

Among all the methods evaluated, *RW-Indiv* turns out best, so we selected this prediction method for constructing the HDN. This outcome of evaluation is based on the following analysis:

- Figure 1 compares the discussed prediction methods for the 20 different diseases shown in Table 3 with respect to the diseases' overall rank. *RW-Indiv* achieves the best overall performance, compared to the other methods.

- For each discussed method M, Table 4 shows the set of diseases for which M produces the best result. *Func-Indiv* and *RW-Indiv* have the largest sets.

- If we comapre the two best candidate methods *Func-Indiv* and *RW-Indiv* to each other according to Figure 1, we conclude that when *Func-Indiv* scores best (e.g., D-6, D-11 and D-19), it is only slightly better than the second-best method; where it is not best, the difference with the best method can be large (e.g., D-3, D-7, D-15, D-16). *RW-Indiv*, on the other hand, never differs much with the best method.

Note that these results are in line with the often observed fact that integrating multiple types of data gives better predictive performance than using one type of data.

3.3. Novel Human Disease Network

Using the *RW-Indiv* prediction method, we have built an HDN for the 20 diseases shown in Table 3. There are 380(20×19) possible edges in the original HDN. Figure 2 shows the score distribution of 380 edges. The X axis shows the 380 edges in the HDN and the Y axis shows the average rank of seed nodes (the smaller, the better). There are two turning points in the curve, roughly at 10% and 90% of all edges. Instead of analyzing the whole HDN, we focus on the pruned HDN containing only the 38 (10% of the original HDN) lowest-weighted edges. Figure 3 shows the pruned HDN in addition to 4 clusters with Graph Density > 0.66. We should emphasize that in our pruned Human Disease Network, each edge shows an informative relationship between two diseases and accordingly, the clustering method should focus on network connections as main indicators of related diseases. We choose the connectivity-based clustering approach using Graph Density, discussed in Schaeffer, 2008 [45] to discover highly-related diseases in the pruned network. The Graph Density of a cluster with N nodes is the ratio of the number of edges E to the number of possible edges, $(\frac{E}{N*(N-1)})$. Compared to other clustering approaches, connectivity-based clustering approaches consider network connections as a main feature in a clustering process, they can be used to filter out noise (outliers), they can discover clusters of arbitrary shape and finally, there is no need to determine the (optimum) number of clusters in advance [46].

For each edge $(d_i) \xrightarrow{rank} (d_j)$, Figure 3 shows the rank of the relationship between two diseases d_i and d_j among all the 380 disease pairs. The highest-ranking found relationship is $(deafness) \xrightarrow{1} (usher\ syndrome)$.

We will now briefly discuss the biological significance of the observed findings. The highest ranked connection (1.) between deafness and Usher's Syndrome is apparent, given the latter is an inherited form of deafness. The link between Deafness and Ehlers-Danlos syndrome however may be attributed also to misdiagnosis of joint laxity, given that the combination of this observation with deafness is more likely to be correctly classified as Stickler syndrome [47]. Epilepsy and Dystonia are both characterized by seizures, and given the proximity of both terms in the Figure also a mechanistic connection between

both disorders can be elucidated. Interesting is the relationship of Long QT Syndrome (LQTS) to both Dystonia and Epilepsy, which hints at the importance of ion channels being important in all of those cases. On the other hand, Anyotrophic Lateral Sclerosis (ALS) is more related to Parkinson's disease (but neither epilepsy nor dystonia), hinting at fundamentally different mechanisms behind those, on the surface similar, disorders characterized by seizures. Apart from this seizure-cluster, also various cancer variations are found to be closely related, namely Xeroderma Pigmentosum (leading to sensitivity to UV light), breast cancer, lymphomas and colorectal cancer. What is interesting is the close link of ALS with the cluster of cancers, since indeed it is assumed that ALS, as a motor neuron disease, may represent a particular case of paraneoplastic encephalomyelitis [48].

3.4. Comparison With Previous Methods

Several studies have focused on the modular nature of human diseases [27, 28, 31, 32]. Goh et al. [27] propose a simple method for building an undirected Human Disease Network. They connect two diseases d_i and d_j in the network if there is at least one gene that is implicated in both. Van Driel et al. [28] calculate phenotype similarities among different disease by applying a text mining approach to the OMIM phenotype records using Medical Subject Headings (MeSH) terms. Linghu et al. [31] construct a gene functional network by combining multiple genomic data sources. Then, they use the functional-linkage network to discover hidden associations between disease pairs having dissimilar phenotypes. More recently, Zhang et al. [32] propose a method much in the same way as Goh et al. [27] that focuses on rare, or orphan, disorders . They call the resulting network "Orphan Disease Network" (ODN).

Complementary to molecular and genetic methods, Hidalgo et al. [49] proposes a Phenotypic Disease Network (PDN) using phenotypic information for human namely, patient clinical histories. In their PDN, nodes are disease phenotypes and edges connect phenotypes that show significant comorbidity. They use Relative Risk [50] and Pearson's correlation to quantify the comorbidities among diseases. In the end, they conclude that the structure of the PDN is relevant to the understanding of disease evolution of patients.

Considering the input information, our proposed method is extensible with respect to considering the information used by the discussed approaches. However Goh et al.'s method [27] is the only one that uses the same input

information as our current input data, and as such is the most natural one to compare to. The main drawback of Goh et al.'s method is in its incapability of discovering relationships among diseases with no common disease-related genes. Our method can be seen as one possible way of solving the limitations of this method. For the comparison, we have applied Goh et al.'s method to our disease dataset, and the resulting HDN is shown on Figure 4. For each edge $(d_i) \leftrightarrow (d_j)$, Figure 4 shows the number of proteins involved in both diseases d_i and d_j ($|SP(d_i) \cap SP(d_j)|$). The best found relationship is $(anemia) \leftrightarrow (hemolytic\ anemia)$. Comparing our proposed HDN (Figure 3) with the disease network discussed by Goh et al. [27] (Figure 4), we observe that our HDN is more informative than the network proposed by Goh et al. [27]: it identifies more relationships between diseases. Obviously, the quality of these relationships still needs to be evaluated; that is done in the next section. Note that, since our method ranks the edges, one can choose to include more or fewer edges in the network as desired; here, we will simply evaluate the HDN obtained with our initial choice of selecting the 10% highest-ranking edges.

3.5. Prediction from Augmented Seed Proteins

We now consider the task of predicting the involvement of proteins in some disease, using a predictive model learned from data. The quality of such a predictive model obviously depends on the prior knowledge incorporated in the data. While earlier models focused on knowledge about one disease, we hypothesize that augmenting the prior knowledge about one disease with knowledge obtained from studying related diseases may yield better predictive models.

To evaluate this hypothesis, we have used our proposed HDN for augmenting the seed proteins of different diseases as follows.

First, we clustered the pruned HDN into n clusters $C_1 \dots C_n$ based on the network connectivity. This gave four clusters, shown in Figure 3. Second, we augmented the seed proteins of each disease by adding the seed proteins of all the diseases in the same cluster C_j; that is, $d_i \in C_j \Rightarrow Aug(d_i) = \cup_{d_k \in C_j} SP(d_k)$, with $Aug(d_i)$ the augmented list of seed proteins of disease d_i. The augmented seed count of each disease is shown in the fourth column of Table 3.

Next, we have evaluated the effect of using the augmented list to learn a model, instead of the original list, along two dimensions: predictive accuracy, as well as biological meaningfulness.

First, we compare the predictions of a model learned using augmented seed proteins (Network approach) with those of a model learned using only the initial seed proteins (Individual approach). In both cases, the same learning method was used (RW-Indiv).

Figure 5 compares the Network approach with the Individual approach for 12 different diseases augmented by our HDN, in terms of the leave-one-out cross validation ranking discussed in section 3.2. The Network approach does not consistently outperform the Individual approach, but can perform much better in individual cases, and does slightly better on average. This suggests that it should not be seen as replacing the Individual approach, but as complementary to it.

Beside the numerical evaluation mentioned above, we also want to evaluate to what extent the proposed network approach predicts biologically meaningful results. To this aim, we try to find literature evidence for proteins predicted to be involved in diseases even when they were not annotated as such in the training data. Our approach consists of the following steps: First, we build a new training set containing the augmented list of seed proteins (positive set) in addition to 100 randomly selected proteins (negative set). Even though we are not sure that all of these random proteins are negative, it is very likely that the majority of them are negative. The remaining positive cases constitute noise in the training set. Second, we build a test set containing all the remaining proteins in the network. Third, we use *RW-Indiv* for predicting new proteins involved in the disease. Tables 5, 6, 7 and 8 show for each of the four clusters shown in Figure 3 the 10 highest-ranked proteins predicted for each cluster.

The first cluster, covering Alzheimer and Ehler-Danlos syndrome, covers both known and potential novel protein targets to treat those diseases. In case of Alzheimer's, BACE2, HSD17B10 and TM2D1 have been implicated in literature before, while COL5A3, which encodes one of the fibrillar collagens, has been established to be involved in Ehler-Danlos syndrome. On the other hand, genes (and proteins) not explicitly associated with those diseases are TGBF2, THBS1 and SPON1, all of which are known to be involved in cell-to-cell interactions, cell-to-matrix interactions, and cell adhesion, respectively. In particular SPON1 can readily be understood to be of importance, given its involvement of attachment of neuron cells and neurite outgrowth.

Similar results covering both established and novel genes are observed for the second cluster, with LQTS, Epilepsy and Dystonia. Dopamine levels and epilepsy have been linked for a long time (DRD1, DRD3 and DRD4; [51]

Dystonia) and they are of practical relevance for treatment. The KCNQ4 ion channel on the other hand has been previously linked with Long QT Syndrom (LQTS). What is interesting, with potential practical implications, is the importance of ALG10 in this cluster, which gates rat ether-a-go-go (the human homolog of the hERG channel involved in LQTS) and which might hence also play an important role in human. No explicit involvement of the EPM2AIP1 gene, encoding laforin, has been described in literature yet; however, our analysis makes a rather strong disease implication for the three diseases present in this cluster.

The third cluster of neoplastic diseases, covering Xeroderma pigmentosum, breast cancer, lymphoma and colorectal cancer gives relatively little surprises, with agreement on MSH3 and MSH6 which are both involved in DNA repair, on the APC tumor suppressor protein, and the RELA oncogene (which bings to the NF kappa b transcription factor with known involvement in cancerogenesis).

The fourth and final disease cluster, of Zellweger syndrome, ALS, and Usher's Syndrome, involves the myosins MYO6, MYO3A and MYO15A which are all known to be involved either in hearing loss or, in the latter case, the actin organization in the hair cells of the cochlea. What is apparent is the link of this set of disorders to the peroxisome, which has been established for this disease cluster before (the involvement of PEX7 and PEX12 which are involved in the assembly of peroxisomes is characteristic, but also ABCD1 is involved in fatty acid transport into the peroxisome, and PXMP3 is involved in its biogenesis). The potentially most surprising gene located in this disease cluster is SIRT3, which is known to be involved in epigenetic silencing and which has been characterized as a potential antineoplastic target - given its prominent role in this analysis, it might hence also play a role for drug treatments of this set of diseases in the future.

Note that our method identifies a particular kind of relationship among diseases that is based on correlations, and there is no guarantee that the newly found gene-disease relationships are causal. Actual causality should be determined experimentally, but the results from the HDN-based analysis can be used to prioritize such experiments. The leave-one-out cross-validation results discussed earlier confirm that the proposed HDN is valuable for such prioritization.

4. Case Study: Long QT Syndrome

In this section, we examine Long QT Syndrome (LQTS) in more detail. According to [52], LQTS is a disorder of the heart's electrical activity which can cause sudden, uncontrollable, dangerous arrhythmias in response to exercise or stress. Table 9 shows the set of proteins involved in LQTS.

4.1. Most Relevant Features for Long QT Syndrome

The number of different functions occurring in our human dataset is 9833; this is also the dimensionality of the *Func-Indiv* method if no dimensionality reduction is used. As we discussed in section 2.3.3, we can use a χ^2-based feature selection methods to reduce this number; at the same time, this technique ranks functions according to how relevant they are for prediction of disease relatedness.

Table 10 shows the ten most discriminant individual functions obtained. It can be seen that the top three GO annotations are explicitly related to cardiac action potential (regulation of heart contraction, regulation of ventricular cardiomyocyte membrane repolarization and negative regulation of sarcomere organization). Positions 4 and 5 are concerning caveolar signaling (which is also very prominent in the heart) and regulation of skeletal muscle contraction, alluding to the fact that muscle contraction in the skeleton and in the heart is goverened by related processes. Membrane rafts (as well as caveolae) are important for cardiac ion channel function as has been found before, [53] which is also correctly identified in Table 10. T-tubule organization, while not immediately apparent, has been linked to a 'new paradigm' for human arrhythmias recently [54]. It is interesting that explicit potassium and ion channel activity are appearing only low in this list, along with the broad term of blood circulation. Hence, overall it can be said that the most discriminative functions are meaningful, with specific functions appearing at the top, biologically derived functions (raft organization, T-tubule organization) in the middle, and general terms at the bottom of the terms derived from the analysis.

Our dataset contains 10,282 proteins. The Anova based method uses the ANOVA measure to select the most relevant among these. This measure has been applied also to the identification of differentially expressed genes in microarray data [55]. More detailed information could be obtained from an ANOVA analysis of the most relevant proteins among the full set of 10,282

proteins. Table 11 now shows the ten proteins with the highest ANOVA measure obtained using our analysis. Interestingly, no ion channel has been most signficiant, but the NADH dehydrogenase NDUFS6. It has been found that HDUSF6 knockouts cause mitochondrial complex I deficieny [56], causing various cardiac problems such as reduced systolic function and cardiac output. On the one hand, this might relate to a functional relationship between diseases; on the other hand it might indicate imperfect diagnosis, hence confusing different underlying disease biology. The six Potassium channels listed can be understood to be involved in direct polarization and depolarization of the cardiac action potential; however the three remaining proteins, namely AKAP6, ALG10B and KCR1 deserve particular attention here. AKAP6 (also called mAKAP) anchors Protein Kinase A to RYR2 which is able to generate Ca2+ 'sparks' due to simultaneous activation within a certain neighborhood radius [57], and hence importance to the cardiac action potential and deviations thereof. ALG10B (also known as KCR1) is interestingly thought to be able to reduce KCNH2 sensitivity to proarrhythmic drug blockade which may be due to glycosylation of this potasssium channel [58, 59], hence our method was able to not only identify protein directly involved in causing LQTS, but also modifier proteins such as AKAP6 and ALG10B.

4.2. Predicting LQTS-Related Proteins using Initial Seed Proteins

In section 3.5, we have already discussed what proteins are predicted as involved in LQTS by the Network based method (Table 6). We here make the same exercise for LQTS using the Individual based approach; in the next section, we will then compare both, illustrating the complementarity between them.

Table 12 lists the highest ranked newly identified LQTS-related genes. In agreement with expectations, many of the genes identified are (as hERG itself) voltage-gated Potassium channels; however also Sodium channels (SNC4A), Calcium channels (CACNB3 and CACNA1A) and solute carriers (SLC8A1) appear in the list. This is in agreement with the known proteins involved in the regulation of cardiac action potential, which are known to involve all three types of ions. KCNJ8 seems to be involved in cardiovascular sudden death at least in mouse models [60], indicating that while focusing on LQTS is of high practical relevance in today's drug development environment, one can in turn also assume that other ion channels involved in drug adverse reactions are currently not receiving sufficient attention. SLC8A1, as a sodium/calcium exchanger, is known to be involved in regulating action potential as well [61],

though it is not easy to find a specific link to the QT interval prolongation in this case. SCN4A mutations have been found to be insignificant under standard conditions, but become relevant in patients treated with LQ-inducing drugs [62]. This finding is interesting since it appears also synergistic adverse relations between genes and LQTS syndrome can be identified using our network approach. One of the potassium channels newly identified to be involved in cardiac action potential regulation (and, hence, with potential LQTS liability) is KCJN12 [63], which is indeed thought to be invovled in providing the cardiac inward rectifier current (IK1). A similar observation can be made regarding KCNA1, where it is thought that a brain-driven cardiac dysfunction can be made responsible for sudden death syndrome in epilespy patients [64]. Mutations in CACNA1 are classified as 'LQTS8' and, while rare, have been shown to be linked to LQTS [65]. Hence, overall we can find associations between the genes identified here and LQTS in many cases - and, interestingly, often they are dependent on the particular genetic or drug treatment conditions of the patients (such as in case of SCN4A and KCNA1).

4.3. Individual vs. Network Based Predictions for LQTS

In this section, we compare the novel LQTS-related proteins predicted by the Network approach and shown in Table 6 with the result of the Individual approach shown in Table 12.

Considerable differences are apparent from the proteins included in the cluster including LQTS along with Epilespy and Dystonia (Table 6), and the prediction of LQTS-related proteins (Table 12). The receptors identified in Table 6 are on the one hand G-Protein Coupled Receptors (GPCRs) such as the Dopamine D1, D3 and D4 receptor subtypes identified with the highest rank in the disease cluster. The only ion channel selected is KCNQ4, which has been linked to deafness [66]; however, only related potassium channels appear to have been linked to LQTS until this stage. On the other hand, KCR1 (ALG10B), which is though to modulate sensitivity to drugs causing LQTS, also appears in this list (as well as in Table 11, in the list of most significant proteins according to ANOVA-based selection). On the other hand, Table 12 is very much dominated by the different subtypes of voltage-gated potassium channels, which occupy 6 out of the 10 positions when RW-indiv is applied to the selection of novel proteins, with the remaining genes selected being ion channels or exchangers of sodium and/or calcium ions. Hence, it can be seen that both methods arrive at a very different selection

of genes involved in the disease cluster, as well as the identification of novel disease genes using the RW-Indiv method. Combined with the fact that very disease relevant genes were identified in Table 12 (as discussed above), we believe that this illustrates the performance of the method implemented in this work.

5. Conclusions and Future Work

The previous studies on analyzing different diseases can be categorized into the Individual and Network approaches. While the Individual approach focuses on one specific disease without considering its relationship with other diseases, the Network approach considers also the diseases relationships.

In this paper, we have proposed, first, a method for building disease networks that considers both functional and structural information in a given PPI network, and, second, a specific Human Disease Network (HDN) resulting from this method. To test the usefulness of this HDN, we have evaluated it in terms of predictive accuracy and biological meaningfulness, and we have compared it to a network constructed using previous methods.

Analyzing different functional and structural prediction methods, we observed that a hybrid method that considers both functional and structural information in the PPI network worked best for building the HDN. We built an HDN for 20 different diseases, and showed that it is biologically meaningful by finding evidence in the literature for the relationships discovered by our method among different diseases. We compared our HDN with one constructed from the same dataset using a previous Network approach [27], and we observed that our method is capable of discovering more (38 versus 8) and still informative relationships among different diseases. In the last step of the evaluation, we clustered the HDN nodes based on their connectivity and we augmented the seed proteins of diseases based on the cluster they belong to. Then, we compared the predictions of models learned using the augmented list with those of models learned using the original list. We observed that considering the relationships among different diseases worked slightly better (9.43 versus 11.01 with respect to average rank of seed proteins) than ignoring them, but, more importantly, it gave highly complementary results both in terms of ranking accuracy and of biological interpretation.

As future work, we could improve and extend the proposed method in several directions. In the first direction, we could apply more extensive validation to the result of our proposed approach. We have already discussed and validated our results in sections 3.3, 3.5 and 4 using literature mining, however, biological/experimental validation of the findings using methods such as PR and RT-PCR is still challenging and needs separate studies. Additionally, although we believe strongly that the results of this paper reduce the search space, generate novel hypothesis and bring new insights for the biologists and clinical researchers, these results should not be assumed as ef-

fective and employed in action by pharmaceutical companies unless they are validated experimentally in the laboratories. In the second direction, in addition to functional and structural feature, we could consider other biological features in the system. In the third direction, we could apply our method to other genomics and metabolomics dataset. We have used the proposed method to augment the initial drug-targets of 200 drugs and a text is being prepared containing these results.

References

[1] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. Brembeck, H. Goehler, M. Stroedicke, Others, A human protein-protein interaction network: A resource for annotating the proteome 122 (2005) 957–968.

[2] T. Milenkovic, N. Przulj, Uncovering biological network function via graphlet degree signatures, Cancer informatics 6 (2008) 257–273.

[3] H. Rahmani, H. Blockeel, A. Bender, Predicting the functions of proteins in PPI networks from global information, in: JMLR Proceeding, 8:82–97, 2010, 2010.

[4] H. N. Chua, W. Sung, L. Wong, Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions, Bioinformatics 22 (2006) 1623–1630.

[5] Z. Lubovac, J. Gamalielsson, B. Olsson, Combining functional and topological properties to identify core modules in protein interaction networks, Proteins 64 (2006) 948–959.

[6] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. Krogan, S. Chung, A. Emili, Others, A bayesian networks approach for predicting protein-protein interactions from genomic data, Science 302 (2003) 449–453.

[7] S. Wong, L. Zhang, A. Tong, Z. Li, D. Goldberg, O. King, G. Lesage, M. Vidal, B. Andrews, H. Bussey, Combining biological networks to predict genetic interactions, Proc Natl Acad Sci USA 101 (2004) 15682–15687.

[8] S. Erten, G. Bebek, R. Ewing, M. Koyutrk, Dada: Degree-aware algorithms for network-based disease gene prioritization., BioData Min 4 (2011).

[9] A. Schlicker, T. Lengauer, M. Albrecht, Improving disease gene prioritization using the semantic similarity of Gene Ontology terms, Bioinformatics 26 (2010) i561–i567.

[10] Z. Dezso, Y. Nikolsky, T. Nikolskaya, J. Miller, D. Cherba, C. Webb, A. Bugrim, Identifying disease-specific genes based on their topological significance in protein networks, BMC Systems Biology 3 (2009) 36+.

[11] S. Köhler, S. Bauer, D. Horn, P. Robinson, Walking the interactome for prioritization of candidate disease genes., American journal of human genetics 82 (2008) 949–958.

[12] L. Sam, Y. Liu, J. Li, C. Friedman, Y. Lussier, Discovery of protein interaction networks shared by diseases, Pac Symp Biocomput (2007) 76–87.

[13] H. Goehler, M. Lalowski, U. Stelzl, S. Waelter, M. Stroedicke, U. Worm, Others, A protein interaction network links git1, an enhancer of huntingtin aggregation, to huntington's disease, Mol Cell 15 (2004) 853–865.

[14] J. Xu, Y. Li, Discovering disease-genes by topological features in human protein-protein interaction network, Bioinformatics 22 (2006) 2800–2805.

[15] M. Oti, B. Snel, M. Huynen, H. Brunner, Predicting disease genes using protein-protein interactions, J Med Genet 43 (2006) 691–698.

[16] X. Wu, S. Li., Cancer Gene Prediction Using a Network Approach. Chapter 11 Mathematical and Computational Biology, Cancer Systems Biology (Ed. Edwin Wang). Series: Chapman and Hall/CRC, 2010.

[17] J. Xu, Y. Li, Discovering disease-genes by topological features in human protein-protein interaction network, Bioinformatics 22 (2006) 2800–2805.

[18] H. Ruffner, A. Bauer, T. Bouwmeester, Human protein-protein interaction networks and the value for drug discovery, Drug Discov Today 12 (2007) 709–716.

[19] V. Neduva, R. Linding, I. Su-Angrand, A. Stark, F. de Masi, T. Gibson, J. Lewis, L. Serrano, R. Russell, Systematic discovery of new recognition

peptides mediating protein interaction networks, PLoS Biol 3 (2005) e405.

[20] A. Ma'ayan, S. Jenkins, J. Goldfarb, R. Iyengar, Network analysis of fda approved drugs and their targets., The Mount Sinai journal of medicine, New York 74 (2007) 27–32.

[21] T. K. B. Gandhi, J. Zhong, S. Mathivanan, L. Karthick, K. N. Chandrika, S. Mohan, S. Sharma, et al., Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets, Nature Genetics 38 (2006) 285–293.

[22] R. Aragues, C. Sander, B. Oliva, Predicting cancer involvement of genes from heterogeneous data, BMC Bioinformatics 9 (2008).

[23] J. Chen, B. Aronow, A. Jegga, Disease candidate gene identification and prioritization using protein interaction networks, BMC Bioinformatics 10 (2009) 73.

[24] T. Milenkovic, V. Memisevic, A. Ganesan, N. Przulj, Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related functional genomics data., Journal of the Royal Society, Interface / the Royal Society 7 (2010) 423–437.

[25] S. Furney, D. Higgins, C. Ouzounis, N. Lopez-Bigas, Structural and functional properties of genes involved in human cancer, BMC Genomics 7 (2006) 3.

[26] L. Li, K. Zhang, J. Lee, S. Cordes, D. Davis, Z. Tang, Discovering cancer genes by integrating network and functional properties, BMC Medical Genomics 2 (2009) 61.

[27] K. Goh, M. Cusick, D. Valle, B. Childs, M. Vidal, A. Barabási, The human disease network, Proceedings of the National Academy of Sciences 104 (2007) 8685–8690.

[28] M. A. van Driel, J. Bruggeman, G. Vriend, H. G. Brunner, J. A. Leunissen, A text-mining analysis of the human phenome., European journal of human genetics : EJHG 14 (2006) 535–542.

[29] M. Oti, H. G. Brunner, The modular nature of genetic diseases, Clinical Genetics 71 (2006) 1–11.

[30] M. Oti, M. A. Huynen, H. G. Brunner, Phenome connections, Trends in Genetics 24 (2008) 103–106.

[31] B. Linghu, E. Snitkin, Z. Hu, Y. Xia, C. DeLisi, Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network, Genome Biology 10 (2009) R91.

[32] M. Zhang, C. Zhu, A. Jacomy, L. J. Lu, A. G. Jegga, The Orphan Disease Networks, American Journal of Human Genetics 88 (2011) 755–766.

[33] H. Rahmani, H. Blockeel, A. Bender, Predicting genes involved in human cancer using network contextual information, J. Integrative Bioinformatics 9 (2012).

[34] S. Berger, A. Ma'ayan, R. Iyengar, Systems Pharmacology of Arrhythmias, Sci. Signal. 3 (2010) ra30+.

[35] R. O. Duda, P. E. Hart, D. G. Stork, Pattern Classification (2nd Edition), Wiley-Interscience, 2000.

[36] P. Domingos, M. Pazzani, On the optimality of the simple bayesian classifier under zero-one loss, Mach. Learn. 29 (1997) 103–130.

[37] H. Liu, R. Setiono, Chi2: Feature selection and discretization of numeric attributes, in: In Proceedings of the Seventh International Conference on Tools with Artificial Intelligence, 1995, pp. 388–391.

[38] H. Rahmani, H. Blockeel, A. Bender, Interaction-based feature selection for predicting cancer-related proteins in protein-protein interaction networks, in: S. Kramer, N. Lawrence (Eds.), Machine Learning in Systems Biology, Proceedings of the Fifth International Workshop, Vienna, Austria, July 20-21, 2011,, 2011, pp. 68–73.

[39] S. Peri, J. D. Navarro, T. Z. Kristiansen, R. Amanchy, V. Surendranath, B. Muthusamy, Others, Human protein reference database as a discovery resource for proteomics., Nucleic Acids Res 32 (2004).

[40] C. Stark, B. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, M. Tyers, Biogrid: a general repository for interaction datasets, Nucleic Acids Research 34 (2006) 535–539.

[41] P. Radivojac, K. Peng, W. Clark, B. Peters, A. Mohan, S. Boyle, S. Mooney, An integrated approach to inferring gene-disease associations in humans., Proteins 72 (2008) 1030–1037.

[42] Go annotation, ???? URL: ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/HUMAN/gene_association.goa_human.gz.

[43] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L. Tranchevent, B. De Moor, P. Marynen, B. Hassan, P. Carmeliet, Y. Moreau, Gene prioritization through genomic data fusion., Nature Biotechnology 24 (2006-05-01 00:00:00.0) 537–44.

[44] T. De Bie, L. Tranchevent, L. M. M. van Oeffelen, Y. Moreau, Kernel-based data fusion for gene prioritization, Bioinformatics 23 (2007) i125–i132.

[45] S. E. Schaeffer, Survey: Graph clustering, Comput. Sci. Rev. 1 (2007) 27–64.

[46] M. Ester, H. P. Kriegel, J. Sander, X. Xu, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, in: E. Simoudis, J. Han, U. Fayyad (Eds.), Second International Conference on Knowledge Discovery and Data Mining, AAAI Press, Portland, Oregon, 1996, pp. 226–231. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.20.2930.

[47] H. Levy, EhlersDanlos syndrome, hypermobility type. In: GeneReviews at GeneTests: Medical Genetics Information Resource (database online). Copyright, University of Washington, Seattle., 2011 (accessed March 5, 2012). http://www.ncbi.nlm.nih.gov/books/NBK1279/.

[48] M. Vigliani, P. Polo, A. Chi, B. Giometto, L. Mazzini, D. Schiffer, Patients with amyotrophic lateral sclerosis and cancer do not differ clinically from patients with sporadic amyotrophic lateral sclerosis, Journal of Neurology 247 (2000) 778–782. 10.1007/s004150070092.

[49] C. A. Hidalgo, N. Blumm, A.-L. Barabasi, N. A. Christakis, A dynamic network approach for the study of human phenotypes, PLoS Comput Biol 5 (2009) e1000353.

[50] C. L. Sistrom, C. W. Garvan, Proportions, odds, and risk, Radiology 230 (2004) 12–19.

[51] M. Starr, The role of dopamine in epilepsy, Synapse 22 (1996) 159–94.

[52] W. I. L. Q. Syndrome?, What Is Long QT Syndrome?, 2011 (accessed October 27, 2011). http://www.nhlbi.nih.gov/health/health-topics/topics/qt/.

[53] A. Maguy, T. E. Hebert, S. Nattel, Involvement of lipid rafts and caveolae in cardiac ion channel function, Cardiovascular Research 69 (2006) 798–807.

[54] M. J. Ackerman, P. J. Mohler, Defining a new paradigm for human arrhythmia syndromes, Circulation Research 107 (2010) 457–465.

[55] I. B. Jeffery, D. G. Higgins, A. C. Culhane, Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data., BMC Bioinformatics 7 (2006) 359.

[56] B.-X. Ke, S. Pepe, D. R. Grubb, J. C. Komen, A. Laskowski, F. A. Rodda, B. M. Hardman, J. J. Pitt, M. T. Ryan, M. Lazarou, J. Kolefff, M. M. H. Cheung, J. J. Smolich, D. R. Thorburn, Tissue-specific splicing of an ndufs6 gene-trap insertion generates a mitochondrial complex i deficiency-specific cardiomyopathy, Proceedings of the National Academy of Sciences (2012).

[57] S.-Q. Wang, L.-S. Song, E. G. Lakatta, H. Cheng, Ca2+ signalling between single l-type ca2+ channels and ryanodine receptors in heart cells (article), Nature 410 (2001) 592–596.

[58] S. Kupershmidt, I. C.-H. Yang, K. Hayashi, J. Wei, S. Chanthaphaychith, C. I. Petersen, D. C. Johns, A. L. George, D. M. Roden, J. R. Balser, The ikr drug response is modulated by kcr1 in transfected cardiac and noncardiac cell lines., FASEB J 17 (2003) 2263–5.

[59] C. I. Petersen, T. R. Mcfarland, S. Z. Stepanovic, P. Yang, D. J. Reiner, K. Hayashi, A. L. George, D. M. Roden, J. H. Thomas, J. R. Balser, In vivo identification of genes that modify ether-a-go-go-related gene activity in Caenorhabditis elegans may also affect human cardiac arrhythmia., Proceedings of the National Academy of Sciences USA 101 (2004) 11773–11778.

[60] G. C. Kane, C.-F. Lam, F. O'Cochlain, D. M. Hodgson, S. Reyes, X.-K. Liu, T. Miki, S. Seino, Z. S. Katusic, A. Terzic, Gene knockout of the kcnj8-encoded kir6.1 katp channel imparts fatal susceptibility to endotoxemia, The FASEB Journal 20 (2006) 2271–2280.

[61] Slc8a1 solute carrier family 8 (sodium/calcium exchanger), member 1 [homo sapiens], ???? URL: `http://www.ncbi.nlm.nih.gov/gene/6546`.

[62] Y. Pereon, G. Lande, S. Demolombe, S. Nguyen The Tich, D. Sternberg, H. Le Marec, A. David, Paramyotonia congenita with an scn4a mutation affecting cardiac repolarization., Neurology 60 (2003) 340–2.

[63] M. Kaibara, K. Ishihara, Y. Doi, H. Hayashi, T. Ehara, K. Taniyama, Identification of human kir2.2 (kcnj12) gene encoding functional inward rectifier potassium channel in both mammalian cells and xenopus oocytes., FEBS Lett 531 (2002) 250–254.

[64] E. Glasscock, J. W. Yoo, T. T. Chen, T. L. Klassen, J. L. Noebels, Kv1.1 potassium channel deficiency reveals brain-driven cardiac dysfunction as a candidate mechanism for sudden unexplained death in epilepsy., The Journal of neuroscience : the official journal of the Society for Neuroscience 30 (2010) 5167–5175.

[65] A. Medeiros-Domingo, P. Iturralde-Torres, M. J. Ackerman, Clinical and genetic characteristics of long qt syndrome (article), Revista Espaola de Cardiologa 60 (2007) 739–752.

[66] P. J. Coucke, P. Van Hauwe, P. M. Kelley, H. Kunst, I. Schatteman, D. Van Velzen, J. Meyers, R. J. Ensink, M. Verstreken, F. Declau, H. Marres, K. Kastury, S. Bhasin, W. T. McGuirt, R. J. H. Smith, C. W. Cremers, P. Van de Heyning, P. J. Willems, S. D. Smith, G. Van Camp, Mutations in the kcnq4 gene are responsible for autosomal dominant

deafness in four dfna2 families, Human Molecular Genetics 8 (1999) 1321–1328.

6. Tables

Table 1: List of the 4 different hybrid methods considering structural and functional information in the network.

Structural Method	Functional Method	Hybrid Method
ST-RW	Func-Indiv	RW-Indiv
ST-RW	Func-Collab	RW-Collab
ST-Anova	Func-Indiv	Anova-Indiv
ST-Anova	Func-Collab	Anova-collab

Table 2: Basic statistical information about our annotated dataset.

Number of Proteins	10,282
Min Degree	1
Max Degree	272
Average Degree	9.39
Number of Proteins with no Function	1519
Average Number of Functions for each Protein	10.40

Table 3: List of the 20 different diseases analyzed in this paper.

Disease ID	Disease Name	Seed Count	Augmented Seed Count
D_1	Alzheimer	7	14
D_2	Amyotrophic	4	17
D_3	Anemia	36	36
D_4	Breast Cancer	21	72
D_5	Cataract	14	14
D_6	Charcot-marie-tooth	11	11
D_7	Colorectal-cancer	20	72
D_8	Deafness	28	28
D_9	Diabetes	23	23
D_10	Dystonia	5	29
D_11	Ehlers-danlos	7	14
D_12	Hemolytic-anemia	11	11
D_13	Epilepsy	11	29
D_14	Long QT Syndrome	13	29
D_15	Lymphoma	27	72
D_16	Mental-retardation	19	19
D_17	Parkinson	8	8
D_18	Usher-syndrome	5	17
D_19	Xeroderma	10	72
D_20	Zellweger	8	17

Table 4: Set of diseases in which each method produces the best result.

Method M	Set of diseases which M produces the best results	Count
ST-RW	D_15, D_16	2
ST-Anova	D_7	1
Func-Indiv	D_2, D_5, D_6, D_8, D_11, D_13, D_19	7
Func-Collab	D_18	1
RW-Indiv	D_1, D_3, D_9, D_10, D_12, D_14, D_20	7
RW-Collab	–	0
Anova-Indiv	D_4, D_17	2
Anova-Collab	–	0

Table 5: 10 highest-ranked proteins predicted for cluster 1 = {Alzheimer, ehler-danlos}.

Index	Protein Symbol	Full Protein Name
1	COL5A3	Collagen, type V, alpha 3
2	THBS1	Thrombospondin 1
3	TGFB2	Transforming growth factor, beta 2
4	COL5A2	Collagen, type V, alpha 2
5	PDGFA	Platelet-derived growth factor alpha polypeptide
6	SPON1	Spondin 1, extracellular matrix protein
7	HSD17B10	Hydroxysteroid (17-beta) dehydrogenase 10
8	HADH2	Hydroxysteroid (17-beta) dehydrogenase 10
9	BACE2	Beta-site APP-cleaving enzyme 2
10	TM2D1	TM2 domain containing 1

Table 6: 10 highest-ranked proteins predicted for each cluster 2 = {LQTS, Epilepsy, Dystonia}.

Index	Protein Symbol	Full protein Name
1	DRD4	Dopamine receptor D4
2	DRD3	Dopamine receptor D3
3	DRD1	Dopamine receptor D1
4	ALG10B	Asparagine-linked glycosylation 10, alpha-1,2-glucosyltransferase homolog B (yeast)
5	KCR1	A membrane Protein That Facilitates Functional Expression of Non-inactivating K+ Currents Associates with Rat EAG Voltage-dependent K+Channels
6	EPM2AIP1	EPM2A (laforin) interacting protein 1
7	KCNQ4	Potassium voltage-gated channel, KQT-like subfamily, member 4
8	TOR1B	Torsin family 1, member B (torsin B)
9	HSPC163	–
10	GCHFR	GTP cyclohydrolase I feedback regulator

Table 7: 10 highest-ranked proteins predicted for cluster 3 = {xeroderma-pigmentosum, breast-cancer-leon, lymphoma, colorectal-cancer}.

Index	Protein Symbol	Protein Full Name
1	MSH6	MutS homolog 6 (E. coli)
2	MSH3	MutS homolog 3 (E. coli)
3	APC	Adenomatous polyposis coli
4	RELA	V-rel reticuloendotheliosis viral oncogene homolog A (avian)
5	TGFBR1	Ransforming growth factor, beta receptor 1
6	PTK2B	PTK2B protein tyrosine kinase 2 beta
7	HIPK2	Homeodomain interacting protein kinase 2
8	RPS6KB1	Ribosomal protein S6 kinase, 70kDa, polypeptide 1
9	TGFB1	Transforming growth factor, beta 1
10	ERBB2	V-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)

Table 8: 10 highest-ranked proteins predicted for cluster 4 = {zellweger-syndrome, amyotrophic-lateral-slecrosis, usher-syndrome}.

Index	Protein Symbol	Protein Full Name
1	MYO15A	Myosin XVA
2	MYO3A	Myosin IIIA
3	MYO6	Myosin VI
4	DDO	D-aspartate oxidase
5	PEX12	Peroxisomal biogenesis factor 12
6	PEX7	Peroxisomal biogenesis factor 7
7	PXMP3	Peroxisomal membrane protein 3
8	SIRT3	Sirtuin 3
9	AGXT	Alanine-glyoxylate aminotransferase
10	ABCD1	ATP-bindende cassette, sub-familie D (ALD), lid 1

Table 9: Proteins associated with the Long QT Syndrome. The data is taken from Berger et. al.[34].

Index	Protein sybmbol	Full Protein name
1	KCNQ1	Potassium voltage-gated channel, KQT-like subfamily, member 1
2	KCNH2	Potassium voltage-gated channel, subfamily H (eag-related), member 2
3	SCN5A	Sodium channel, voltage-gated, type V, alpha subunit
4	ANK2	Ankyrin 2, neurona l
5	KCNE1	Potassium voltage-gated channel, Isk-related family, member 1
6	KCNE2	Potassium voltage-gated channel, Isk-related family, member 2
7	KCNJ2	Potassium inwardly-rectifying channel, subfamily J, member 2
8	CACNA1C	Calcium channel, voltage-dependent, L type, alpha 1C subunit
9	CAV3	Caveolin 3
10	SCN4B	Sodium channel, voltage-gated, type IV, beta
11	AKAP9	A kinase (PRKA) anchor protein (yotiao) 9
12	SNTA1	Syntrophin, alpha 1
13	ALG10	Asparagine-linked glycosylation 10 homolog (yeast, alpha-1,2-glucosyltransferase)

Table 10: 10 most discriminative functions according to $\chi^2(f_i)$ (Formula 5).

Index	Function	Short Description
1	GO:0008016	Regulation of heart contraction
2	GO:0060307	Regulation of ventricular cardiomyocyte membrane repolarization
3	GO:0060299	Negative regulation of sarcomere organization
4	GO:0002095	Caveolar macromolecular signaling complex
5	GO:0014819	Regulation of skeletal muscle contraction
6	GO:0031579	Membrane raft organization
7	GO:0033292	T-tubule organization
8	GO:0005251	Delayed rectifier potassium channel activity
9	GO:0005244	Voltage-gated ion channel activity
10	GO:0008015	Blood circulation

Table 11: 10 most discriminative proteins according to Anova (Formula 3).

Index	Protein	Short Description
1	NDUFS6	NADH dehydrogenase [ubiquinone] iron-sulfur protein 6, mitochondrial
2	KCNH1	Potassium voltage-gated channel subfamily H member 1
3	KCNH5	Potassium voltage-gated channel, subfamily H (eag-related), member 5
4	KCNF1	Potassium voltage-gated channel subfamily F member 1
5	AKAP6	A-kinase anchor protein 6
6	ALG10B	Asparagine-linked glycosylation 10, alpha-1,2-glucosyltransferase homolog B
7	KCR1	A membrane Protein That Facilitates Functional Expression of Non-inactivating K+ Currents Associates with Rat EAG Voltage-dependent K+Channels
8	KCNE1	Potassium voltage-gated channel subfamily E member 1
9	KCNH2	potassium voltage-gated channel, subfamily H (eag-related), member 2
10	KCNE2	Potassium voltage-gated channel subfamily E member 2

Table 12: Newly identified LQTS-related proteins by applying *RW-Indiv* method to the original seed proteins.

index	Gene-Name	Short Description
1	KCNH1	Potassium voltage-gated channel, subfamily H (eag-related), member 1
2	KCNH5	Potassium voltage-gated channel, subfamily H (eag-related), member 5
3	KCNJ8	Potassium inwardly-rectifying channel, subfamily J, member 8
4	SLC8A1	Solute carrier family 8 (sodium/calcium exchanger), member 1
5	SCN4A	Sodium channel, voltage-gated, type IV, alpha subunit
6	KCNJ4	Potassium inwardly-rectifying channel, subfamily J, member 4
7	CACNB3	Calcium channel, voltage-dependent, beta 3 subunit
8	KCNJ12	Potassium inwardly-rectifying channel, subfamily J, member 12
9	KCNA1	potassium voltage-gated channel, shaker-related subfamily, member 1 (episodic ataxia with myokymia)
10	CACNA1A	Calcium channel, voltage-dependent, P/Q type, alpha 1A subunit

7. Figures Captions

Figure 1: Average rank of seed proteins in 20 different diseases shown in Table 3. *RW-Indiv* achieves the best overall performance comparing to the other methods and is therefore a good candidate method for building the HDN.

Figure 2: The score distribution of the 380 edges in HDN. The X axis shows the 380 edges in the HDN and the Y axis shows the average rank of seed nodes (the smaller, the better). There are two turning points $x = 38$ and $x = 331$ in this figure.

Figure 3: Pruned Human Disease Network by keeping only 38 (10% of the original HDN) highest-ranked relationships among different diseases. In this network, each node d_i is a disease and each edge $(d_i) \xrightarrow{rank} (d_j)$ is a rank of the relationship between two diseases d_i and d_j among all the 380 disease pairs. The best found relationship is $(deafness) \xrightarrow{1} (usher\ syndrome)$. We cluster the pruned HDN based on the network connectivity and the 4 discovered clusters are shown in grey squares in this figure. As it can be seen, the different cancer types present in the dataset share the same cluster (cluster #3), while other clusters (e.g. cluster #4) share diseases that are also known to exhibit related phenotypes (such as Usher syndrome and Zellweger syndrome, which are both frequently characterized by visual and hearing impairments, among other effects).

Figure 4: Human Disease Network based on the common proteins (Proposed by Goh et al. [27]). In this network, each node d_i is a disease and each edge $(d_i) \leftrightarrow (d_j)$ is the number of common proteins between two related diseases.

Figure 5: Comparing the Network approach with the Individual approach for 12 different diseases augmented by our HDN. For several diseases, the Network based approach outperforms the Individual approach with respect to ranking the seed proteins in a leave-one-out cross validation.

8. Figures

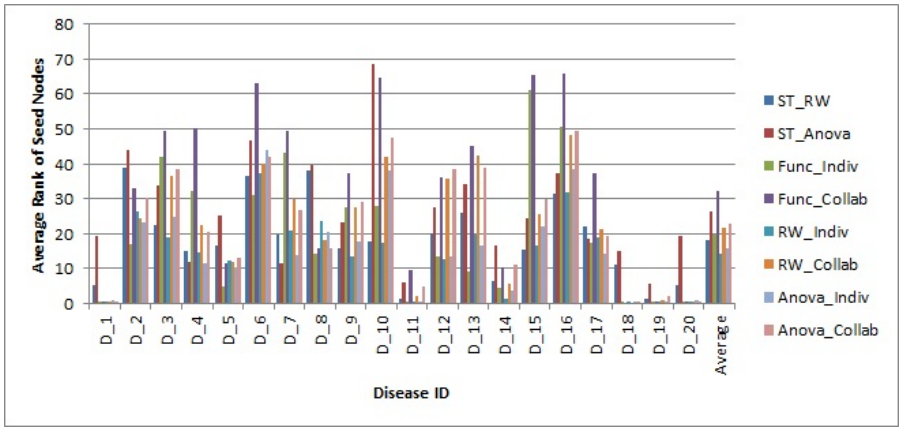

Figure 1: Average rank of seed proteins in 20 different diseases shown in Table 3. *RW-Indiv* achieves the best overall performance comparing to the other methods and is therefore a good candidate method for building the HDN.

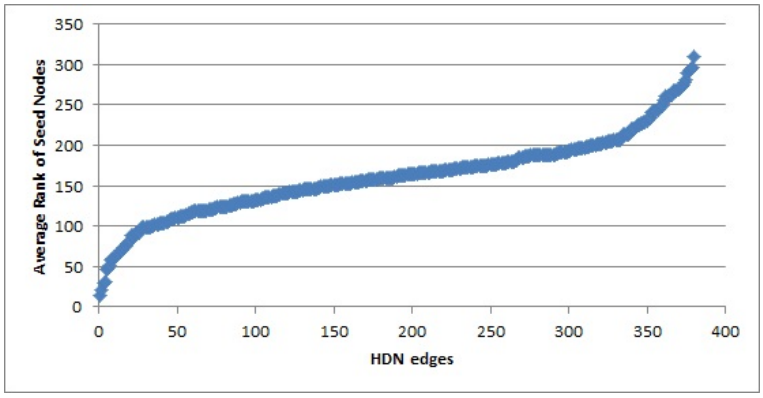

Figure 2: The score distribution of the 380 edges in HDN. The X axis shows the 380 edges in the HDN and the Y axis shows the average rank of seed nodes (the smaller, the better). There are two turning points $x = 38$ and $x = 331$ in this figure.

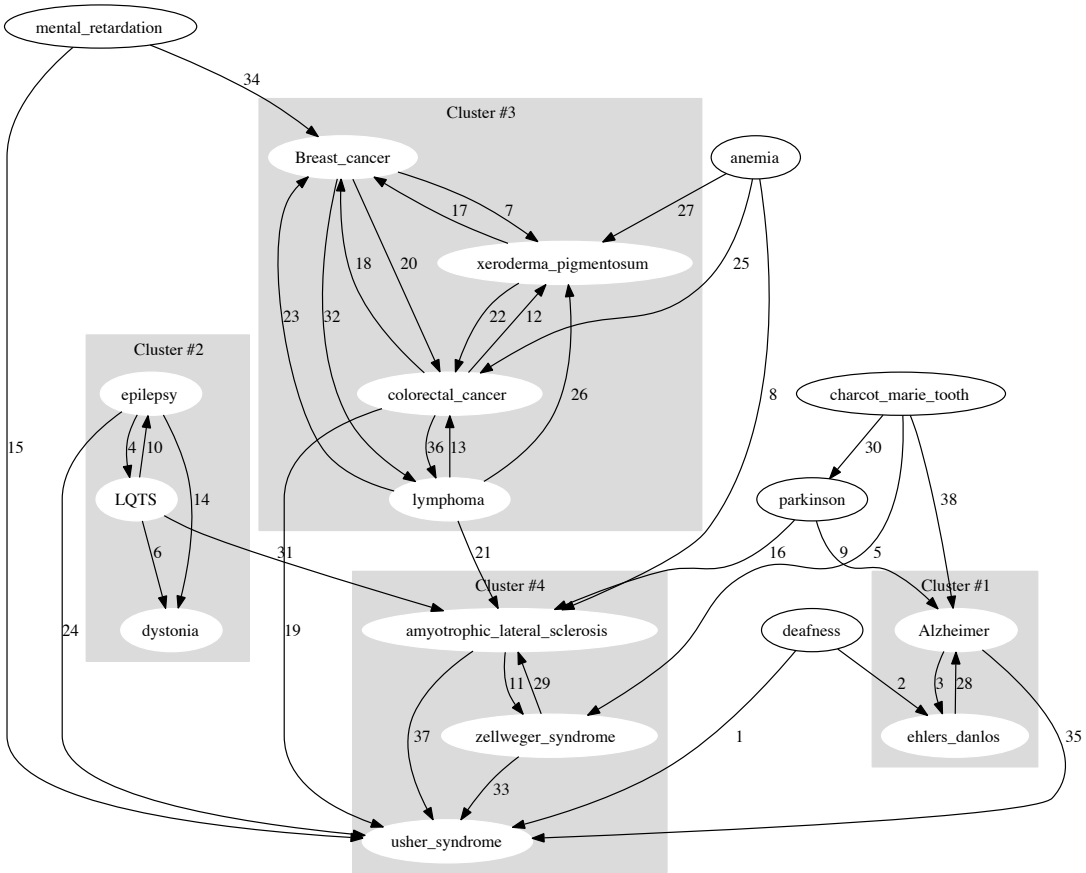

Figure 3: Pruned Human Disease Network by keeping only 38 (10% of the original HDN) highest-ranked relationships among different diseases. In this network, each node d_i is a disease and each edge $(d_i) \xrightarrow{rank} (d_j)$ is a rank of the relationship between two diseases d_i and d_j among all the 380 disease pairs. The best found relationship is $(deafness) \xrightarrow{1} (usher\ syndrome)$. We cluster the pruned HDN based on the network connectivity and the 4 discovered clusters are shown in grey squares in this figure. As it can be seen, the different cancer types present in the dataset share the same cluster (cluster #3), while other clusters (e.g. cluster #4) share diseases that are also known to exhibit related phenotypes (such as Usher syndrome and Zellweger syndrome, which are both frequently characterized by visual and hearing impairments, among other effects).

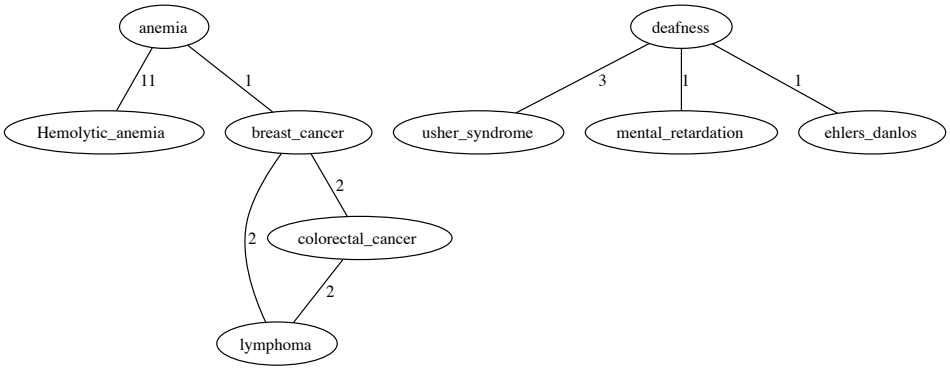

Figure 4: Human Disease Network based on the common proteins (Proposed by Goh et al. [27]). In this network, each node d_i is a disease and each edge $(d_i) \leftrightarrow (d_j)$ is the number of common proteins between two related diseases.

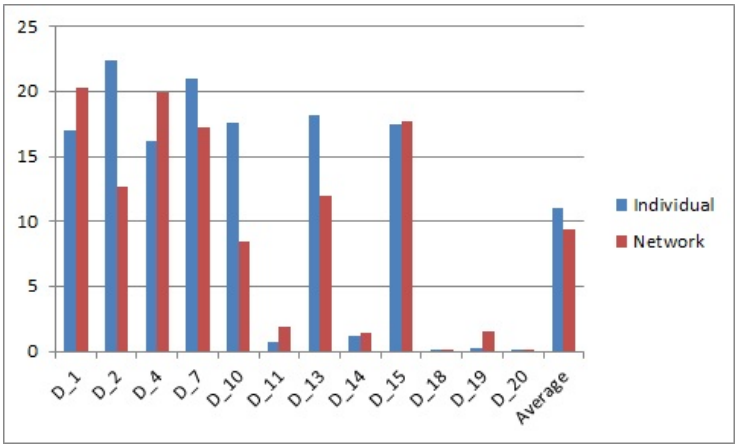

Figure 5: Comparing the Network approach with the Individual approach for 12 different diseases augmented by our HDN. For several diseases, the Network based approach outperforms the Individual approach with respect to ranking the seed proteins in a leave-one-out cross validation.